Multiscale Support Vector Learning With Projection Operator Wavelet Kernel for Nonlinear Dynamical System Identification

Zhao Lu, *Senior Member, IEEE*, Jing Sun, *Fellow, IEEE*, and Kenneth Butts, *Member, IEEE*

***Abstract*— A giant leap has been made in the past couple of decades with the introduction of kernel-based learning as a mainstay for designing effective nonlinear computational learning algorithms. In view of the geometric interpretation of conditional expectation and the ubiquity of multiscale characteristics in highly complex nonlinear dynamic systems [1]–[3], this paper presents a new orthogonal projection operator wavelet kernel, aiming at developing an efficient computational learning approach for nonlinear dynamical system identification. In the framework of multiresolution analysis, the proposed projection operator wavelet kernel can fulfill the multiscale, multidimensional learning to estimate complex dependencies. The special advantage of the projection operator wavelet kernel developed in this paper lies in the fact that it has a closed-form expression, which greatly facilitates its application in kernel learning. To the best of our knowledge, it is the first closed-form orthogonal projection wavelet kernel reported in the literature. It provides a link between grid-based wavelets and mesh-free kernel-based methods. Simulation studies for identifying the parallel models of two benchmark nonlinear dynamical systems confirm its superiority in model accuracy and sparsity.**

***Index Terms*— Composite kernel, linear programming support vector regression (LP-SVR), multiscale modeling, nonlinear systems identification, orthogonal projection operator, raised-cosine wavelet.**

I. Introduction

KERNEL-BASED support vector (SV) learning was originally proposed for solving nonlinear classification and recognition problems, and it marked the beginning of a new era in the computational learning from examples paradigm [4]–[7]. Thereafter, the rationale of SV learning has been successfully generalized to various fields such as nonlinear regression, signal processing, integral equation, and path planning [8]–[11]. When SV learning is employed for function approximation and estimation, the approaches are often referred to as the SV regression (SVR). As a typical nonparametric kernel learning approach, SVR also provides a promising avenue to nonlinear dynamical systems modeling.

Recently, a new line of research has been initiated for developing novel nonstandard kernels for meshless methods in solving the partial differential equations in the realm of computational mathematics [12]–[14]. Even the term *kernel engineering* has been coined lately, because efficient algorithms require specially tailored application-dependent kernels [14]. On the other hand, although the past decade has witnessed intensive research activities on the kernel-based computational learning methods, most of the researchers use standard kernel functions, such as Gaussian radial basis function (RBF) kernel and polynomial kernel. It was pointed out that the kernel machine with the widely used Gaussian RBF kernel is endowed with the capability little more than a template matcher, and some inherited drawbacks, such as the locality of the kernel function, may result in the inefficiency in representing highly complex functions by kernel expansion [15]. Hence, one objective of this paper is to bridge this gap by exploring the computational capability of the nonstandard kernel in kernel-based SV learning: the closed-form orthogonal wavelet is exploited to construct a multiscale projection operator wavelet kernel for complex systems modeling and prediction.

It has been revealed in recent studies that in modeling highly complex nonlinear dynamical systems, the multiscale SV learning is more capable and flexible over conventional single-scale SV learning [14]. In particular, in [14], it was emphasized that the success and failure of kernel usage may crucially depend on proper scaling. As the keystone of nonlinear SV learning, the construction of kernel functions plays an important role in fulfilling the efficacious multiscale SV learning. For identifying nonlinear dynamical systems, the wavelet outperforms the (windowed) Fourier transform due to its aptitude in capturing very short-lived high-frequency phenomena, such as transients in signals [16]. Albeit some efforts have been made to develop wavelet kernel functions for SV learning [17]–[21], weaving multiresolution wavelet analysis into modern kernel learning is not a trivial task, because almost all known orthonormal wavelets, except for the Haar and the Shannon, cannot be expressed in the closed form in terms of elementary analytical functions, such as the trigonometric, exponential, or rational functions [22], [23]. Discontinuities in the Haar wavelet and poor time localization of the Shannon wavelet have limited their applicability in multiscale modeling problems [23].

Manuscript received January 18, 2015; revised December 10, 2015; accepted December 20, 2015. Date of publication January 5, 2016; date of current version December 22, 2016. This work was supported by Toyota Motor Engineering and Manufacturing North America, Inc.

Z. Lu is with the Department of Electrical Engineering, Tuskegee University, Tuskegee, AL 36088 USA (e-mail: zlu@ieee.org).

J. Sun is with the Department of Naval Architecture and Marine Engineering and the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jingsun@umich.edu).

K. Butts is with Toyota Motor Engineering and Manufacturing North America, Ann Arbor, MI 48105 USA (e-mail: ken.butts@tema.toyota.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2513902

The closed-form representation of the Morlet wavelet was employed for constructing a single-scale kernel function in [17] and [18], but the lack of interscale orthogonality and intrascale orthogonality makes it difficult to be used for implementing the multiscale kernel learning in a systematic way. In addition, different from the anisotropic stationary multiscale wavelet kernel in [24], where the symmetry requirement is relaxed and the kernel proposed is translation (but no longer rotational) invariant, the kernel function developed herein can be viewed as a class of novel finite expansion kernels [25], which is not translation invariant. Inspired by the geometric purport of conditional expectation [1]–[3], novel projection operator wavelet kernels are developed for linear programming SV learning, and it excels in multiscale learning and modeling for complex nonlinear dynamical systems.

In the realm of nonlinear dynamic system identification, the nonlinear autoregression with exogenous (NARX) input model is widely used for representing discrete-time nonlinear systems, and the regressor for the NARX model consists of two parts: 1) an autoregressive (AR) part and 2) a moving-average (MA) part. The mathematical description of the NARX model is as follows:

$$\hat{y}_n = f(y_{n-1}, y_{n-2}, \ldots, y_{n-P}, u_n, u_{n-1}, \ldots, u_{n-Q+1}) \quad (1)$$

where u_n and y_n are the input and output to the system at time instant t_n, and the vectors $\boldsymbol{y}_{n-1} = [y_{n-1}, y_{n-2}, \ldots, y_{n-P}]^T$ and $\boldsymbol{u}_n = [u_n, u_{n-1}, \ldots, u_{n-Q+1}]^T$ are the AR and MA parts, respectively. The AR part is a window of past system outputs with output order P, and the MA part is a window of past and current system inputs with input order Q. The NARX model (1) is also called the series–parallel model, because the system and model are parallel with respect to u_n but in series with respect to y_n.

Essentially, the identification of the NARX model can be formulated as a nonlinear function regression problem. It amounts to modeling the conditional expectation of system output y_n, given the regression vector consisting of the AR part $\boldsymbol{y}_{n-1}$ and the MA part $\boldsymbol{u}_n$, i.e., $E[y_n|\boldsymbol{y}_{n-1}, \boldsymbol{u}_n]$ [26]. From the geometric standpoint from the linear operator theory, the conditional expectations implement the projections onto linear subspaces as best approximation in essence [2]. This enlightens us to conceive innovative wavelet-based projection operator kernels for multiscale SV learning. In this paper, by integrating the multiresolution wavelet analysis and kernel learning systems, a new computational learning approach to nonlinear system identification and predictions is developed. To confirm and validate the effectiveness of the proposed learning strategy, the identification of parallel models of nonlinear dynamical systems is used as a touchstone for the simulation study. Contrary to the series–parallel model (1), where the past values of the system input and the system output constitute the regressor, the regressor of the parallel model is composed of the past values of the system input and the model output, that is

$$\hat{y}_n = f(\hat{y}_{n-1}, \hat{y}_{n-2}, \ldots, \hat{y}_{n-P}, u_n, u_{n-1}, \ldots, u_{n-Q+1}). \quad (2)$$

Model (2) can be simulated as standalone without using the real data as inputs. It is, however, well known that the identification of a parallel model is much more challenging than that for a series–parallel model due to the feedback involved in the model [27], [28].

The following generic notations will be used throughout this paper: non-boldface symbols, such as $x, y, \alpha, \ldots$, refer to scalar valued objects, lower case boldface symbols, such as $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\beta}, \ldots$, refer to vector valued objects, and capital boldface symbols, such as $\mathbf{K_1}, \mathbf{K_2}, \ldots$, will be used for matrices.

II. Closed-Form Orthogonal Wavelet in Multiresolution Analysis

Multiresolution analysis is conceptualized by a coarse-to-fine sequence of embedded closed linear subspaces $\{V_j\}_{j\in Z} \subseteq L^2(R)$ as follows.

Definition 1: A multiresolution analysis is a decomposition of $L^2(R)$ into a chain of nested subspaces $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots V_{j-1} \subset V_j \subset V_{j+1} \cdots$ such that the following holds.

1) (Separation) $\cap_{j\in Z} V_j = V_{-\infty} = \{0\}$.
2) (Density) $\overline{\cup_{j\in Z} V_j} = V_\infty = L^2(R)$.
3) (Self-similarity in scale) $f(x) \in V_0$ if and only if $f(2^j x) \in V_j$, $j \in Z$.
4) There exists a scaling function $\varphi \in V_0$ whose integer-translates span the space V_0, and for which the set $\{\varphi(x-k), k \in Z\}$ is an orthonormal basis.

Here, j is the index of resolution level. The function φ is called the scaling function, since its dilates and translates constitute orthonormal bases for all approximation subspaces V_j, and the orthogonal complement of V_j in V_{j+1}, i.e., the direct difference $W_j = V_{j+1} \ominus V_j$, is called the wavelet space or detail space. ◊

By successively decomposing the approximation spaces as $V_{j+1} = V_j \oplus W_j$, where $\oplus$ denotes the orthogonal direct sum, the functional space $L^2(R)$ can be decomposed as an orthogonal direct sum of wavelet spaces of different resolutions, i.e., $\overline{\oplus_{j\in Z} W_j} = L^2(R)$. The wavelet function ψ can be defined, such that $\{\psi(x-k)\}_{k\in Z}$ is an orthonormal basis of W_0, and W_j is the span of orthonormal wavelet functions $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, i.e., $W_j = \text{span}(\{\psi_{j,k}\}_{k\in Z})$. Obviously, the scaling functions and a wavelet are orthogonal whenever the scaling functions are of lower resolution [29].

Almost all the known orthonormal wavelets, except for the Haar wavelet and the Shannon wavelet, cannot be expressed in the closed form or in terms of simple analytical functions. Instead, they can only be expressed as the limit of a sequence or the integral of some functions [22], [23]. This has been a main stumbling block to developing wavelet kernels for multiscale kernel learning and modeling. In this paper, the type-II raised-cosine wavelet, a recently discovered closed-form orthonormal wavelet family [22], [30], will be capitalized on to develop innovative and effective projection operator wavelet kernels with multiscale and spatially varying resolution properties for SV learning. Its advantages will be demonstrated by identifying the parallel models of two benchmark nonlinear dynamical systems.

As in signal reconstruction technology, the raised- cosine scaling function is derived from its power spectrum

(energy spectrum), defined as follows [22], [30]:

$$|\hat{\varphi}(\omega)|^2 = \begin{cases} 1, & |\omega| \le \pi(1-b) \\ \dfrac{1}{2}\left[1+\cos\dfrac{|\omega|-\pi(1-b)}{2b}\right] & \\ & \pi(1-b) \le |\omega| \le \pi(1+b) \\ 0, & |\omega| \ge \pi(1+b) \end{cases} \tag{3}$$

where $\hat{\varphi}(\omega)$ is the Fourier transform of scaling function $\varphi(x)$, i.e., $\hat{\varphi}(\omega) = \int_{-\infty}^{\infty} \varphi(t) e^{-i\omega t} dt$. From (3), it follows that the spectrum of the scaling function involves the positive and complex square roots:

$$\hat{\varphi}_1(\omega) = \begin{cases} 1, & 0 \le |\omega| \le \pi(1-b) \\ \cos\left[\dfrac{|\omega|-\pi(1-b)}{4b}\right] & \\ & \pi(1-b) \le |\omega| \le \pi(1+b) \\ 0, & |\omega| \ge \pi(1+b) \end{cases} \tag{4}$$

$$\hat{\varphi}_2(\omega) = \begin{cases} 1, & 0 \le \omega \le \pi(1-b) \\ \dfrac{1}{2}\left[1+e^{i(1/2b)(\omega-\pi(1-b))}\right] & \\ & \pi(1-b) \le \omega \le \pi(1+b) \\ 0, & \omega \ge \pi(1+b). \end{cases} \tag{5}$$

The scaling functions can be found using the inverse Fourier transform as follows:

$$\varphi_1(x) = \frac{\sin\pi(1-b)x + 4bx\cos\pi(1+b)x}{\pi x(1-(4bx)^2)} \tag{6}$$

$$\varphi_2(x) = \frac{\sin\pi(1-b)x + \sin\pi(1+b)x}{2\pi x(1+2bx)} = \frac{\cos(\pi bx)}{(1+2bx)}\operatorname{sinc}(\pi x). \tag{7}$$

The scaling functions $\varphi_1(x)$, $\varphi_2(x)$ correspond to the type-I and type-II raised-cosine wavelets. In this paper, the type-II raised-cosine wavelet, derived from the scaling function $\varphi_2(x)$, is our primary concern. To derive the type-II raised-cosine wavelet function $\psi_2(x)$ from the explicit form of $\varphi_2(x)$, one may apply Theorem 1 directly [30].

Theorem 1: Let $\wp$ be the set of all $g \in L^1(R)$, such that $g(x) \ge 0$ and supp $g \subset [-\pi/3, \pi/3]$, and then $g(x)$ is even; $\int_{-v}^{v} g(x)dx = \pi$ for some $0 < v \le \pi/3$ where supp $g = \overline{\{x \in R | g(x) \ne 0\}}$. For each $g \in \wp$, the function $\varphi(x)$ defined by its spectrum

$$\hat{\varphi}(\omega) = \frac{1}{2} + \frac{1}{2}\exp i\vartheta(\omega) \tag{8}$$

where $\vartheta(\omega) = \int_{-\omega-\pi}^{\omega-\pi} g(x)dx$ is a real band-limited orthonormal cardinal scaling function, and the corresponding mother wavelet function $\psi(x)$ is given by

$$\psi(x) = 2\varphi(2x-1) - \varphi\left(\frac{1}{2}-x\right). \tag{9}$$

$\Diamond$

The rigorous proof of this theorem can be found in [30]. Evidently, the type-II raised-cosine scaling function spectrum $\hat{\varphi}_2(\omega)$ given by (5) is in the form of (8). Hence, it follows from

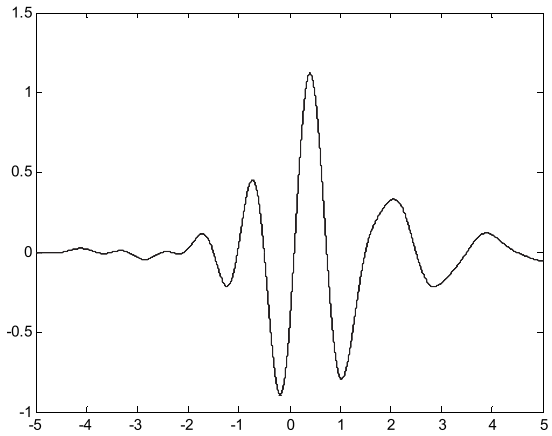

Fig. 1. Type-II raised-cosine wavelet function.

Theorem 1 that the type-II raised-cosine wavelet function is in the form of [22]:

$$\begin{aligned} \psi_2\left(x+\frac{1}{2}\right) &= \frac{1}{2\pi x[1+4bx]}[\sin 2\pi(1-b)x + \sin 2\pi(1+b)x] \\ &\quad - \frac{1}{2\pi x[1-2bx]}[\sin\pi(1-b)x + \sin\pi(1+b)x] \\ &= \frac{2\cos(2\pi bx)}{1+4bx}\operatorname{sinc}(2\pi x) - \frac{\cos(\pi bx)}{1-2bx}\operatorname{sinc}(\pi x) \end{aligned} \tag{10}$$

where the sinc function, closely related to the spherical Bessel function of the first kind, is defined as $\operatorname{sinc}(x) = \sin x/x$. Parallel to the scaling functions, the raised-cosine wavelet functions $\psi_1(x)$ and $\psi_2(x)$ are both band-limited functions, and the type-II raised-cosine wavelet function (10) is plotted in Fig. 1.

As eigenfunctions of the Calderón–Zygmund operator [31], orthogonal wavelets have exceptional potential for modeling high-dimensional, multiscaled input–output maps. The mother wavelet ψ gives birth to an entire family of wavelets by means of two operations: 1) dyadic dilations and 2) integer translations. Let j denote the dilation index and k represent the translation index, and each wavelet born of the mother wavelet is indexed by both of these indices

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{11}$$

for integer-valued j and k. A wavelet (which, when appropriately dilated, forms the basis for the detail spaces) must be localized in time, in the sense that $\psi(x) \to 0$ quickly as $|x|$ gets large [29]. Similarly, the family of scaling functions (father wavelets) takes the form of

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k). \tag{12}$$

In the literature, constructing the stationary (translation-invariant) wavelet kernels by defining $k(x, y) = \psi(x-y)$ is a popular approach for multiscale learning [17], [18], [20], [24]. In contrast to that, the multiscale wavelet kernels are constructed based on the notion of the orthogonal projection operator in Section III.

III. Projection Operator Wavelet Kernel

A projection P on an inner product space X is known as orthogonal projection or orthogonal projector if the image of P and the null space of P are orthogonal, i.e., if $\mathrm{Im}(P) \perp \mathrm{Null}(P)$ [32], [33]. In multiresolution analysis, every $f \in L^2(R)$ can be approximated arbitrarily accurately by its orthogonal projections $\mathrm{proj}_{V_{j+1}} f$ on the approximation space V_{j+1}, and then the orthogonal projection of f on V_{j+1} can be decomposed as the summation of orthogonal projections on V_j and W_j [16]

$$\mathrm{proj}_{V_{j+1}} f = \mathrm{proj}_{V_j} f + \mathrm{proj}_{W_j} f. \tag{13}$$

The complement $\mathrm{proj}_{W_j} f$ provides the details of f that appear at the scale j but disappear at the coarser scale, and a given function $f \in L^2(R)$ can be decomposed as the summation of its projection on the wavelet spaces W_j as follows:

$$f = \sum_j \mathrm{proj}_{W_j} f = \sum_j \sum_k c_{j,k} \psi_{j,k} = \sum_j \sum_k c_{j,k} \psi(2^j x - k) \tag{14}$$

which is called the wavelet series expansion.

By defining the integral operator kernel function $Q(x, y) = \sum_k \varphi(x-k)\varphi(y-k)$, the orthogonal projection operator $E_j : L^2(R) \to V_j$ can be written in terms of kernel function $Q(x, y)$

$$\begin{aligned} E_j[f] = \mathrm{proj}_{V_j} f(x) &= \sum_k \left(\int \varphi_{jk}(y) f(y) dy \right) \varphi_{jk}(x) \\ &= \int 2^j Q(2^j x, 2^j y) f(y) dy. \end{aligned} \tag{15}$$

By defining

$$Q_j(x, y) = \sum_k \varphi_{j,k}(x) \varphi_{j,k}(y) = 2^j \sum_k \varphi(2^j x - k) \varphi(2^j y - k) \tag{16}$$

the orthogonal projection operator E_j can be simplified as the integral operator with kernel $Q_j(x, y)$, that is

$$E_j[f] = \mathrm{proj}_{V_j} f(x) = \int Q_j(x, y) f(y) dy. \tag{17}$$

Analogous to the integral operator representation of $\mathrm{proj}_{V_j} f(x)$, the projection $\mathrm{proj}_{W_j} f(x)$ of f on wavelet space can also be represented in the form of an integral operator. To this end, Theorem 2 is needed.

Theorem 2 [32], [34]: Let V_{j+1} and V_j be the closed linear subspace of $L^2(R)$, and let E_{j+1} and E_j be the orthogonal projections onto V_{j+1} and V_j, respectively. The difference $D_j = E_{j+1} - E_j$ is an orthogonal projection if and only if $V_j \subset V_{j+1}$. The range of D_j is $W_j = V_{j+1} \ominus V_j$, which is the orthogonal complement of V_j in V_{j+1}. ◊

For the difference between the projection operator E_{j+1} and E_j, i.e., $D_j = E_{j+1} - E_j$, it follows from Theorem 2 that D_j is also an orthogonal projection operator on the subspace W_j and it can be precisely given by:

$$\begin{aligned} D_j[f] = \mathrm{proj}_{W_j} f(x) &= \sum_k \left(\int \psi_{jk}(y) f(y) dy \right) \psi_{jk}(x) \\ &= \int 2^j K(2^j x, 2^j y) f(y) dy \end{aligned} \tag{18}$$

where $K(x, y) = \sum_k \psi(x-k)\psi(y-k)$. If one defines

$$K_j(x, y) = \sum_k \psi_{j,k}(x) \psi_{j,k}(y) = 2^j \sum_k \psi(2^j x - k) \psi(2^j y - k) \tag{19}$$

the orthogonal projection operator D_j can be represented by the following integral operator with kernel $K_j(x, y)$:

$$D_j[f] = \mathrm{proj}_{W_j} f(x) = \int K_j(x, y) f(y) dy. \tag{20}$$

As the kernel of the projection integral operator onto W_j, $K_j(x, y)$ is called an orthogonal projection operator wavelet kernel. Note that it is based on the rigorous multiresolution analysis framework and has an analytic expression in terms of a raised-cosine wavelet function, thereby enabling multiscale learning by dyadic dilation.

In the presence of irregular localized features, multiresolution learning algorithms are necessary to take local as well as global complexities of the input–output map into account. In this way, underfitting and overfitting can be avoided simultaneously in approximating highly nonlinear functions [35]. In effect, multiresolution approximation is a mathematical process of hierarchically decomposing the input–output approximation to capture both the macroscopic and microscopic features of the system behavior [36]. The unknown function underlying any given measured input–output data can be considered as consisting of high-frequency local input–output variation details superimposed on the comparatively low-frequency smooth background. At each stage, finer details are added to the coarser description, providing a successively better approximation to the input–output data. The multiscale learning strategy developed herein aims to take advantage of the multiresolution structure of wavelets to provide spatially varying resolution, and for this purpose, the orthogonal projection operator wavelet kernel (22) can be extended to multiscale kernels [25], [37], [38], according to Theorems 3 and 4.

Theorem 3 [33], [34]: Let $W_j, W_{j+1}, \ldots, W_{j+m}$ be a family of closed linear subspaces of $L^2(R)$, and let $D_j, D_{j+1}, \ldots, D_{j+m}$ be the orthogonal projections on $W_j, W_{j+1}, \ldots, W_{j+m}$, respectively. The finite sum of orthogonal projection operators

$$\Theta = D_j + D_{j+1} + \cdots + D_{j+m} \tag{21}$$

is an orthogonal projection operator if and only if the subspaces W_{j+k} $(k = 0, 1, \ldots, m)$ are pairwise orthogonal. In this case, the range of operator Θ is $W_j \oplus W_{j+1} \oplus \cdots \oplus W_{j+m}$. ◊

Hence, the multiscale orthogonal projection operator wavelet kernel can be constructed as the summation of $K_j(x, y)$ as follows:

$$\widetilde{K}(x, y) = \sum_j K_j(x, y) = \sum_j \sum_k \psi_{j,k}(x) \psi_{j,k}(y) \tag{22}$$

and it follows from Theorem 3 that the integral operator with kernel (22) fulfills the orthogonal projection onto the direct sum of wavelet subspaces at different scales $W_{j_{\min}} \oplus W_{j_{\min}+1} \oplus \cdots \oplus W_{j_{\max}}$. The developed kernel (22) shares a similar form to the finite multiscale kernel proposed in [25],

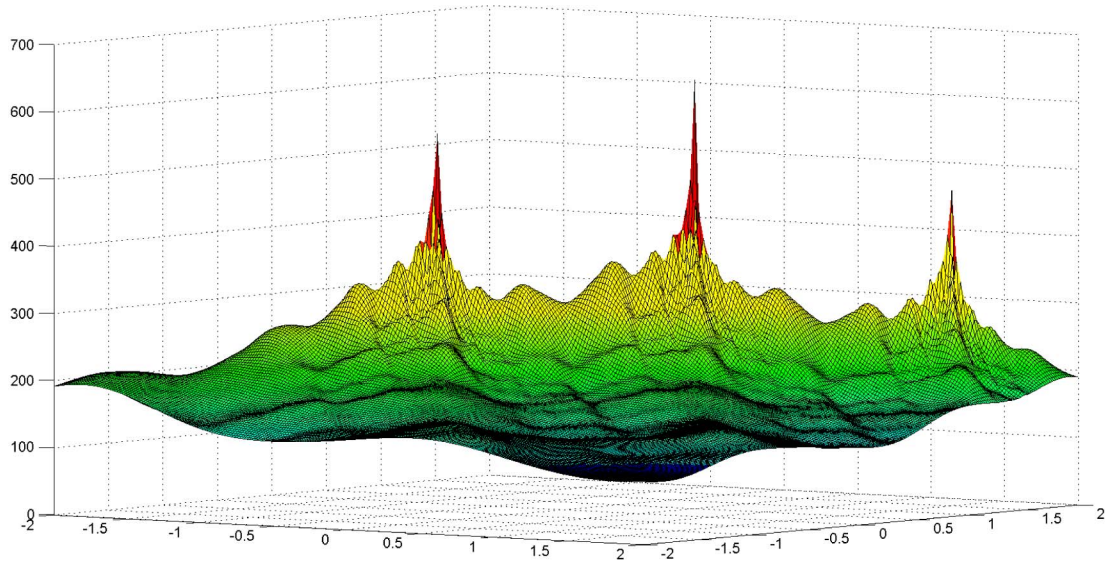

Fig. 2. 2-D multiscale orthogonal projection operator wavelet kernel (24).

which is also called the expansion kernel in [12]. However, the construction of finite multiscale kernels in [25] is based on the superposition of shifts and scales of a single compactly supported function on grids. In contrast, the raised-cosine wavelet function used in (22) is a band-limited wavelet [39], which is impossible to be compactly supported in the time-domain according to the well-known duration-bandwidth theorem (uncertainty principle) [40]. Nevertheless, the raised-cosine wavelet function is fast decaying in time (contrary to the poor time localization of the Shannon wavelet), which also makes the evaluation of the kernel inexpensive.

The addition of multiple single-scale wavelet kernels at different scales provides multiscale wavelet kernels with more flexibility than single-scale wavelet kernels [37], [38]. For a given system, the range of level index j in kernel (22) needs to be tailored to the particular application. Furthermore, for estimating the multivariate dependencies, the multiscale projection operator wavelet kernels need to be extended to the multidimensional space. Based on the fact that a wavelet basis in higher dimensions can be obtained by taking the tensor product of 1-D wavelet bases, the construction of a multidimensional wavelet kernel function can be carried out using Theorem 4 [41], [42].

Theorem 4: Let a multidimensional set of functions be defined by the basis functions that are the tensor products of the coordinatewise basis functions. Then, the kernel that defines the inner product in the n-dimensional basis is the product of n 1-D kernels. $\Diamond$

Hence, as the inner product in high-dimensional space, the multidimensional multiscale orthogonal projection operator wavelet kernel can be constructed as the product of 1-D multiscale kernels

$$\widetilde{K}(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{d} \sum_{j} \lambda_j \sum_{k} \psi_{j,k}(x_i)\psi_{j,k}(y_i) \tag{23}$$

where d is the dimension, and $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^T$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_d]^T$. To illustrate the multiscale characteristics of the developed closed-form projection operator wavelet kernel, an exemplary 2-D projection operator wavelet kernel given by

$$\widetilde{K}(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{2} \sum_{j=-7}^{7} \lambda_j \sum_{k=-1}^{1} \psi_{j,k}(x_i)\psi_{j,k}(y_i) \tag{24}$$

is plotted in Fig. 2.

IV. Liner Programming SVR With Composite Kernel

In quadratic programming SVR (QP-SVR), the smoothness is used as a prior for regularizing the function to ensure the generalization, i.e., the smooth functions having few or small variations are more likely. However, recent research indicates that the smoothness prior alone could be problematic and insufficient in learning highly nonlinear functions with many steep and/or smooth variations, which characterize the kind of complex task needed for artificial intelligence (AI) [15]. Hence, instead of using the smoothness prior as that in quadratic programming SV learning, linear programming SVR (LP-SVR) takes an entirely different avenue to build the model by the regularization technique.

A model identified through the SVR is represented as the kernel expansion on the SVs, which are the data points in a selected subset of the training data [4]–[6]. In other words, the model is represented in a data-dependent nonparametric form. In the endeavor of applying kernel learning strategies for identifying nonlinear dynamical systems, the idea of the composite kernel was conceptualized and developed for taking into account the different cause–effect relationships of the AR and MA parts to the NARX model output instead of assimilating them [43], [44]. The model represented by a composite kernel expansion is in the form of

$$\hat{y}_n = \sum_{i=1}^{N} \beta_i (k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) \tag{25}$$

where β_i is the expansion coefficient and N is the number of sampled data. k_1 and k_2 are the kernel functions for the AR and MA parts respectively, and $k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)$ is defined as the composite kernel. The composite kernel expansion model (25) enables us to use different kernel functions for the AR and MA parts of the regressor in (1).

The vector pairs $[(\boldsymbol{y}_{i-1})^T, (\boldsymbol{u}_i)^T]^T$ corresponding to the nonzero coefficients β_i in model representation (25) are the SVs. Consequently, the model sparsity, which is defined as the ratio of the number of SVs to the number of all training data points, plays a critical role in controlling model complexity and alleviating model redundancy. A kernel expansion model with substantial redundant terms is against the parsimonious principle that ensures the simplest possible model that explains the data, and may deteriorate the generalization performance and increase the computational requirements substantially.

The number of nonzero components in the coefficient vector $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_N]^T$ largely determines the complexity of the kernel expansion model (25). In order to enforce the sparseness of the model, the linear programming SV learning is considered here, instead of QP-SVR. It employs the ℓ_1 norm of the coefficient vector $\boldsymbol{\beta}$ in model (25) as a regularizer in the objective function to control the model complexity and structural risk. By introducing the ε-insensitive loss function, which is defined as

$$L(y_n - \hat{y}_n) = \begin{cases} 0, & \text{if } |y_n - \hat{y}_n| \le \varepsilon \\ |y_n - \hat{y}_n| - \varepsilon, & \text{otherwise} \end{cases} \tag{26}$$

the regularization problem to be solved becomes

$$\min\ R_{\text{reg}}[f] = \|\boldsymbol{\beta}\|_1 + C \sum_{n=1}^{N} L(y_n - \hat{y}_n) \tag{27}$$

where the parameter C controls the extent to which the regularization term influences the solution and ε is the error tolerance. Geometrically, the ε-insensitive loss function defines a ε-tube. The idea of using the ℓ_1 norm to secure a sparse representation is also explored in the emerging theory of compressive sensing [45], [46].

By introducing the slack variables ξ_n, $n = 1, 2, \ldots, N$ to accommodate otherwise infeasible constraints and to enhance robustness, the regularization problem (27) can be transformed into the following equivalent constrained optimization problem:

$$\min\ \|\boldsymbol{\beta}\|_1 + C \sum_{n=1}^{N} \xi_n$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{N} \beta_i (k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) - y_n \le \varepsilon + \xi_n \\ y_n - \sum_{i=1}^{N} \beta_i (k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) \le \varepsilon + \xi_n \\ \xi_n \ge 0, \quad n = 1, 2, \ldots, N \end{cases} \tag{28}$$

where the constant $C > 0$ determines the tradeoff between the sparsity of the model and the amount up to which deviations larger than ε can be tolerated. For the purpose of converting (28) into a linear programming problem, the components β_i of the coefficient vector $\boldsymbol{\beta}$ and their absolute values $|\beta_i|$ are decomposed as follows:

$$\beta_i = \alpha_i^+ - \alpha_i^- \qquad |\beta_i| = \alpha_i^+ + \alpha_i^- \tag{29}$$

where $\alpha_i^+, \alpha_i^- \ge 0$, and for a given β_i, there is an unique pair (α_i^+, α_i^-) fulfilling both the equations in (29). Note that both the variables cannot be positive at the same time, i.e., $\alpha_i^+ \cdot \alpha_i^- = 0$. In this way, the optimization problem (28) can be reformulated as

$$\min \sum_{i=1}^{N} (\alpha_i^+ + \alpha_i^-) + C \sum_{n=1}^{N} \xi_n$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-)(k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) - \xi_n \le \varepsilon + y_n \\ -\sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-)(k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) - \xi_n \le \varepsilon - y_n \\ \xi_n \ge 0, \quad n = 1, 2, \ldots, N. \end{cases} \tag{30}$$

Next, define the vector

$$\boldsymbol{c} = (\underbrace{1, 1, \ldots, 1}_{N}, \underbrace{1, 1, \ldots, 1}_{N}, \underbrace{C, C, \ldots, C}_{N})^T \tag{31}$$

and write the ℓ_1 norm of $\boldsymbol{\beta}$ as

$$\|\boldsymbol{\beta}\|_1 = (\underbrace{1, 1, \ldots, 1}_{N}, \underbrace{1, 1, \ldots, 1}_{N}) \begin{pmatrix} \boldsymbol{\alpha}^+ \\ \boldsymbol{\alpha}^- \end{pmatrix} \tag{32}$$

with the N-dimensional column vectors $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ defined as $\boldsymbol{\alpha}^+ = (\alpha_1^+, \alpha_2^+, \ldots, \alpha_N^+)^T$ and $\boldsymbol{\alpha}^- = (\alpha_1^-, \alpha_2^-, \ldots, \alpha_N^-)^T$, and the constrained optimization problem (30) can be cast as a linear programming problem in the following form:

$$\min\ \boldsymbol{c}^T \begin{pmatrix} \boldsymbol{\alpha}^+ \\ \boldsymbol{\alpha}^- \\ \boldsymbol{\xi} \end{pmatrix}$$
$$\text{s.t.} \begin{cases} \begin{pmatrix} \mathbf{K_1} + \mathbf{K_2} & -(\mathbf{K_1} + \mathbf{K_2}) & -\mathbf{I} \\ -(\mathbf{K_1} + \mathbf{K_2}) & \mathbf{K_1} + \mathbf{K_2} & -\mathbf{I} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\alpha}^+ \\ \boldsymbol{\alpha}^- \\ \boldsymbol{\xi} \end{pmatrix} \le \begin{pmatrix} \boldsymbol{y} + \varepsilon \\ \varepsilon - \boldsymbol{y} \end{pmatrix} \\ \boldsymbol{\alpha}^+, \quad \boldsymbol{\alpha}^- \ge 0, \quad \boldsymbol{\xi} \ge 0 \end{cases} \tag{33}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_N)^T$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^T$, and $\mathbf{I}$ is an $N \times N$ identity matrix. $\mathbf{K_1}$ and $\mathbf{K_2}$ are the kernel matrices with entries defined as $(\mathbf{K_1})_{in} = k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1})$ and $(\mathbf{K_2})_{in} = k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)$. The calculation of the vectors $\boldsymbol{\alpha}^+$, $\boldsymbol{\alpha}^-$ and the SVs selection can be accomplished by solving the linear optimization problem (33) using the well-known simplex algorithm or the primal–dual interior point algorithm. With the solution to linear programming problem (33), the coefficients of the composite kernel expansion model (25) can be calculated using (29), and thereby model (25) can be built as follows:

$$\hat{y}_n = \sum_{i \in \text{SV}} \beta_i (k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)). \tag{34}$$

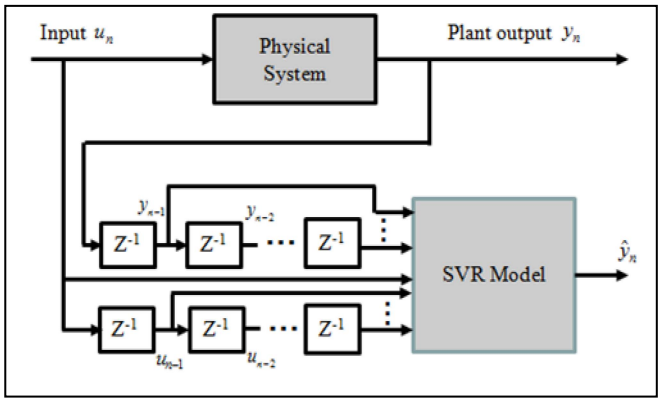

Fig. 3. Model in series–parallel configuration.

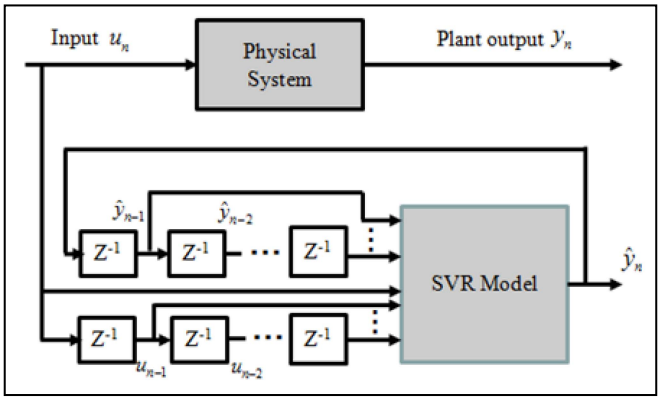

Fig. 4. Model in parallel configuration.

This composite kernel expansion on the selected SVs is for representing the nonlinear dynamics underlying the time series $\{u_i, y_i\}$, $i = 1, 2, \ldots, N$. Contrary to the QP-SVR where all the data points not inside the ε-tube are selected as SVs, the LP-SVR still can generate a sparse solution even when the ε is set to be zero.

Most of the preceding works applying SV learning to nonlinear system identification treat system identification as a general regression problem, where the AR and MA parts are consolidated in the regressor [47]–[50]. However, the chosen single kernel function might be ineffective in characterizing different cause–effect relationships of the AR and MA parts to the model output. Modeling the different dependencies by heterogeneous kernel functions is the main motivation for using the composite kernel, which provides new degrees of freedom in representing nonlinear dynamics. The use of composite kernel also makes the model more amenable to control law design, which also blazes a new path to control-oriented sparse modeling.

V. Application to Nonlinear Dynamical System Identification

Machine learning is centered around making predictions, based on already identified trends and properties in the training data set. Forecasting future behavior based on past observations has also been a long standing topic in system identification and time series analysis. According to the different regression vectors used, the identification model for nonlinear dynamical systems can be categorized as the series–parallel model (or NARX model), and parallel model (or nonlinear output error model) [27], [28]. For the series–parallel model (see Fig. 3), the past values of the input and output of the actual system form the regression vector in order to produce the estimated output $\hat{y}_n$ at time instant t_n. For the parallel model depicted in Fig. 4, however, the regression vector is composed of the past values of the input and output of the identification model. Thus, without coupling to the real systems, the parallel models are emancipated from relying on the outputs of the actual systems. In effect, the parallel model is a recurrent NARX model, whose computational capability is equivalent to a Turing machine [51], [52]. The identification of the series–parallel model amounts to building a one-step ahead predictor, while the identification of the parallel model is for the long-term prediction.

In the realm of nonlinear systems identification, there is general consensus that one of the most formidable technical challenges is building a model usable in parallel configuration, which is much more intractable than building a series–parallel model due to the feedback involved [27], [28]. However, a multitude of applications, e.g., fault detection and diagnosis, predictive control, simulation, and so on, require a parallel model, since the prediction of many steps into the future is needed.

In theory, long-term predictions can be obtained from a short-term predictor, for example, a one-step ahead predictor, simply by applying the short predictor many times (steps) in an iterative way. This is called iterative prediction, and lays the foundation for obtaining a parallel model by training in the series–parallel configuration [53]–[55]. Another way called direct prediction provides a once-completed predictor with a long-term prediction step, and the specified multistep prediction can then be obtained directly from the established predictor in a manner similar to computing one-step predictions [53], [54]. The main downside of the direct modeling approach is that it requires different models for different steps ahead prediction. It is generally believed that the iterative prediction approach is in most cases more efficient than the direct approach assuming that the dynamics underlying the time series are correctly specified by the model [53].

In this simulation study, to demonstrate the superiority and effectiveness of the proposed novel kernel function for nonlinear dynamical systems modeling, the LP-SVR learning algorithm with multiscale orthogonal projection operator wavelet kernel is used to build parallel models for the benchmark hydraulic robot arm data set and Box and Jenkins' data set in the spirit of the iterative prediction approach. Although these data sets have been widely used for the performance evaluation of various system identification methods in [17], [48], [49], and [56]–[58], most of the work reported in the literature focuses on the identification of the series–parallel models and their parallel models have rarely been studied.

Partitioning the benchmark data sets into training and validation subsets, the identification procedure includes two phases. First, the one-step ahead predictor, i.e., the

series–parallel model, is identified on the training data set in the series–parallel configuration as that in Fig. 3, and then in the second phase, the attained one-step ahead predictor is iteratively used in parallel configuration for the long-term prediction on the validation data set, as shown in Fig. 4.

For the sake of comparison, several commonly used kernel functions are employed for modeling on the same data sets as well, such as the Gaussian RBF kernel defined by

$$k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2}\right) \tag{35}$$

the polynomial kernel defined by

$$k(\boldsymbol{x}, \boldsymbol{z}) = (1 + \langle \boldsymbol{x}, \boldsymbol{z} \rangle)^q \tag{36}$$

the inverse multiquadric kernel defined by

$$k(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{\sqrt{\|\boldsymbol{x} - \boldsymbol{z}\|^2 + c^2}} \tag{37}$$

the B-spline kernel defined by

$$k(\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{d} B_{2J+1}(x_i - z_i) \tag{38}$$

and the Morlet wavelet kernel given by

$$k(\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{d} \phi\left(\frac{x_i - z_i}{\delta}\right) \tag{39}$$

where $\phi(x) = \cos(1.75x)\exp(-x^2/2)$ and δ, σ, q, c, and J are all the adjustable parameters of the above kernel functions. For the B-spline kernel, B-spline function $B_\ell(\cdot)$ represents a particular example of a convolutional basis and can be expressed explicitly as [59]

$$B_\ell(x) = \frac{1}{\ell!}\sum_{r=0}^{\ell+1}\binom{\ell+1}{r}(-1)^r\left(x + \frac{\ell+1}{2} - r\right)_+^\ell \tag{40}$$

where the function $(\cdot)_+$ is defined as the truncated power function, that is

$$x_+ = \begin{cases} x, & \text{for } x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

A. Hydraulic Robot Arm Dynamical System Identification

For the hydraulic robot arm dynamical system, the position of a robot arm is controlled by a hydraulic actuator. The control input u_n represents the size of the valve opening through which oil flows into the actuator, and the output y_n is a measure of the oil pressure that determines the robot arm position. In modeling this dynamical system, for fair comparison, the same regressor $[y_{n-1}, y_{n-2}, y_{n-3}, u_{n-1}, u_{n-2}]$ and the data set partition scheme as those in the literature [17], [48], [49], [56] are adopted herein. The first half of the data set containing 511 training data pairs is used for training in series–parallel configuration, and the other half for validation data in parallel configuration.

In the training phase, model (34) with $\boldsymbol{y}_{n-1} = [y_{n-1}, y_{n-2}, y_{n-3}]$ and $\boldsymbol{u}_n = [u_{n-1}, u_{n-2}]$ is learned by LP-SVR to attain the one-step ahead approximator. Upon training completion, our objective is to provide satisfactory multistep prediction without using the actual system output y_n, i.e., to validate the model in parallel configuration as follows:

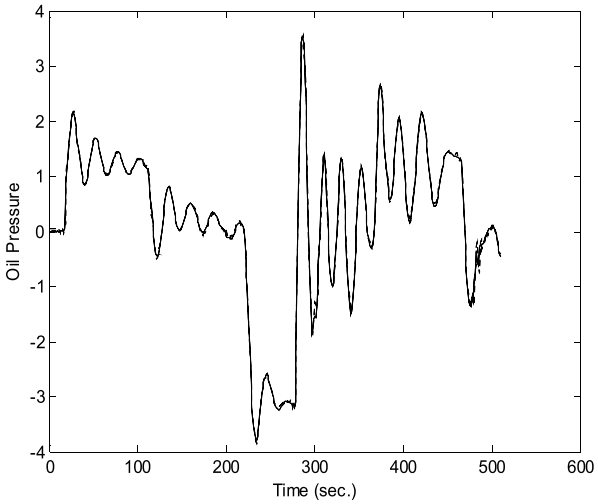

Fig. 5. Training in series–parallel configuration for model (34) of robot arm by LP-SVR with projection operator wavelet kernel (44) and (45). Solid line: actual system output. Dotted line: model output.

$$\hat{y}_n = \sum_{i \in \mathrm{SV}} \beta_i (k_1(\boldsymbol{y}_{i-1}, \hat{\boldsymbol{y}}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) \tag{42}$$

where $\hat{\boldsymbol{y}}_{n-1} = [\hat{y}_{n-1}, \hat{y}_{n-2}, \hat{y}_{n-3}]$. The approximation accuracies on the training and validation data sets are evaluated by calculating the root mean square error (RMSE)

$$E_{\mathrm{rms}} = \sqrt{\frac{1}{M}\sum_{n=1}^{M}[\hat{y}_n - y_n]^2} \tag{43}$$

where $\hat{y}_n$ is the estimated output of the model and M is the number of data in the data set for evaluation. The validation accuracy is crucial in assessing the generalization performance of the model. In applying SVR with kernel functions to train the model, manual tuning of the kernel parameters as well as ε and C for optimum results is required.

The parameters used for learning are $\varepsilon = 0.06$ and $C = 5$, and the projection operator wavelet kernels $k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1})$ and $k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)$ in the composite kernel expansion model take the following forms respectively:

$$\widetilde{K}_1(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=-9}^{-4}\Psi_j(x_1, z_1)\sum_{j=-6}^{-3}\Psi_j(x_2, z_2)\sum_{j=-12}^{-3}\Psi_j(x_3, z_3) \tag{44}$$

$$\widetilde{K}_2(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=-10}^{5}\Phi_j(x_1, z_1)\sum_{j=-2}^{2}\Phi_j(x_2, z_2) \tag{45}$$

where $\Psi_j(x_i, z_i) = (1/2^j)\sum_{k=-10}^{10}\psi_{j,k}(x_i)\psi_{j,k}(z_i)$ with the kernel parameter 0.0002, and $\Phi_j(x_i, z_i) = (1/2^j)\sum_{k=-6}^{6}\varphi_{j,k}(x_i)\varphi_{j,k}(z_i)$ with the kernel parameter 0.001. The training result based on the multiscale projection operator wavelet kernel (44) and (45) is illustrated in Fig. 5, and the training RMSE is 0.0745. The attained model is subsequently

TABLE I
ROBOT ARM PARALLEL MODEL IDENTIFICATION BY LP-SVR WITH DIFFERENT COMPOSITE KERNEL FUNCTIONS

Kernel functions k_1 and k_2	SV ratio	Training RMSE	Validation RMSE
Gaussian RBF kernel	11.0%	0.2005	0.8456
Polynomial kernel	2.0%	0.0913	0.7089
B-Spline kernel	16.4%	0.0717	0.4940
Inverse multi-quadric kernel	5.5%	0.1189	0.7167
Morlet wavelet kernel	3.1%	0.1450	0.6816
Projection operator wavelet kernel	13.9%	0.0745	0.4112

TABLE II
ROBOT ARM PARALLEL MODEL IDENTIFICATION BY QP-SVR WITH DIFFERENT COMPOSITE KERNEL FUNCTIONS

Kernel functions k_1 and k_2	SV ratio	Training RMSE	Validation RMSE
Gaussian RBF kernel	40.7%	0.3024	0.6973
Polynomial kernel	27.6%	0.0893	0.7386
B-Spline kernel	31.7%	0.0529	0.6296
Inverse multi-quadric kernel	50.9%	0.1483	0.6532
Morlet wavelet kernel	32.9%	0.1493	0.7365

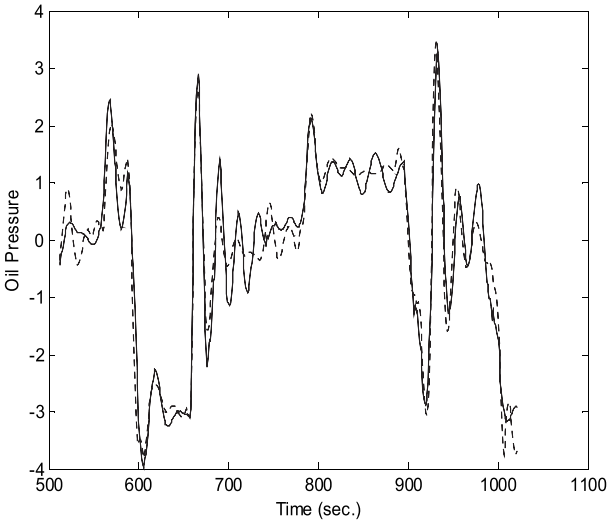

Fig. 6. Validation in parallel configuration for mode (42) of robot arm by LP-SVR with projection operator wavelet kernel (44) and (45). Solid line: actual system output. Dotted line: model output.

validated on the validation data set in parallel configuration for long-term/mid-term prediction, and the validation result is shown in Fig. 6 and Table I.

Following the same procedure, the kernel functions (35)–(39) are also applied to train model (34) by LP-SVR and QP-SVR. After tuning the parameters for optimum results, the RMSEs on the training and validation data sets together with model sparsity obtained by these comparative models are listed in Table I for LP-SVR and Table II for QP-SVR. The corresponding plots can be found in [24].

Measured by the SV ratio, the sparsity of the model with the closed-form projection operator wavelet kernel is commensurate with the models adopting other kernel functions; moreover, it is very evident that the projection operator wavelet kernel considerably outperforms other kernel functions in terms of validation accuracy in parallel configuration, which implies excellent generalization performance.

In parallel configuration, the errors for the sth-step prediction are the accumulation of the errors of the previous $(s-1)$ steps. In general, the longer the forecasting horizon, the larger the accumulated errors are and the less accurate the iterative method is. Hence, it is remarkable that, while using the identical regressor on the same training and validation data sets, this parallel model validation accuracy is even better than some of those obtained in the series–parallel configuration by

TABLE III
GAS FURNACE PARALLEL MODEL IDENTIFICATION BY LP-SVR WITH DIFFERENT COMPOSITE KERNEL FUNCTIONS

Kernel functions k_1 and k_2	SV ratio	Training RMSE	Validation RMSE
Gaussian RBF kernel	1.3%	1.1853	1.9001
Polynomial kernel	12.8%	0.7204	2.7905
B-Spline kernel	73.8%	0.0988	3.2156
Inverse multi-quadric kernel	1.3%	1.3254	2.6664
Morlet wavelet kernel	68.5%	0.2094	2.6469
Projection operator wavelet kernel	16.1%	0.2298	0.5148

TABLE IV
GAS FURNACE PARALLEL MODEL IDENTIFICATION BY QP-SVR WITH DIFFERENT COMPOSITE KERNEL FUNCTIONS

Kernel functions k_1 and k_2	SV ratio	Training RMSE	Validation RMSE
Gaussian RBF kernel	80.5%	1.0421	2.6201
Polynomial kernel	77.9%	0.6667	2.4571
B-Spline kernel	62.4%	0.8426	1.9846
Inverse multi-quadric kernel	78.5%	0.6140	2.4761
Morlet wavelet kernel	88.6%	0.1109	2.7435

other popular learning strategies. For example, the RMSE was 0.467 for a one-hidden-layer sigmoid neural network case and 0.579 for a wavelet network case [56].

In terms of computing time for training, LP-SVR is around seven times faster than QP-SVR on this data set (Intel Core i5 processor), and the computing resource required by QP-SVR might become prohibitively expensive when increasing the size of the training data set. It is also notable when comparing the model sparsities in Tables I and II that the LP-SVR substantially exceeds the QP-SVR in producing succinct model representations.

B. Box and Jenkins' Identification Problem

The Box and Jenkins' gas furnace data set was recorded from a combustion process of a methane-air mixture. The original data set consists of 296 input–output data pairs that were recorded at a sampling rate of 9 s. The gas combustion process has one input variable, gas flow rate u_n, and one output variable, the concentration of carbon dioxide (CO_2) in the outlet gas, y_n. The instantaneous value of the output y_n can be regarded as being influenced by ten variables $y_{n-1}, y_{n-2}, \ldots, y_{n-4}$ and $u_{n-1}, u_{n-2}, \ldots, u_{n-6}$ [57], [58]. In modeling this dynamical system, the regressor $[y_{n-1}, y_{n-2}, u_{n-2}, u_{n-3}, u_{n-4}]$ is employed herein. The first 150 data pairs are used for training in the series–parallel configuration, and the subsequent 90 data pairs are used for validation in parallel configuration. Due to the different distribution and magnitude order of the measurements in the data set, proper data rescaling is necessary.

In training model (34) with $\boldsymbol{y}_{n-1} = [y_{n-1}, y_{n-2}]$ and $\boldsymbol{u}_n = [u_{n-2}, u_{n-3}, u_{n-4}]$, the kernel functions $k_1(\boldsymbol{y}_{i-1}, \boldsymbol{y}_{n-1})$ and $k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)$ take the following forms, respectively:

$$\widetilde{K}_1(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=-9}^{2} \Psi_j(x_1, z_1) \sum_{j=1}^{4} \Psi_j(x_2, z_2) \tag{46}$$

$$\widetilde{K}_2(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=0}^{2} \Psi_j(x_1, z_1) \sum_{j=2}^{3} \Psi_j(x_2, z_2) \sum_{j=1}^{2} \Psi_j(x_3, z_3) \tag{47}$$

where $\Psi_j(x_i, z_i) = \sum_{k=-2}^{0} \psi_{j,k}(x_i) \psi_{j,k}(z_i)$ and the kernel parameters are 1.22 and 0.013, respectively. The training result based on the multiscale projection operator wavelet kernel (46) and (47) is illustrated in Fig. 7, and the corresponding RMSE is 0.2298. Subsequently, the model is validated in parallel configuration

$$\hat{y}_n = \sum_{i \in \mathrm{SV}} \beta_i (k_1(\boldsymbol{y}_{i-1}, \hat{\boldsymbol{y}}_{n-1}) + k_2(\boldsymbol{u}_i, \boldsymbol{u}_n)) \tag{48}$$

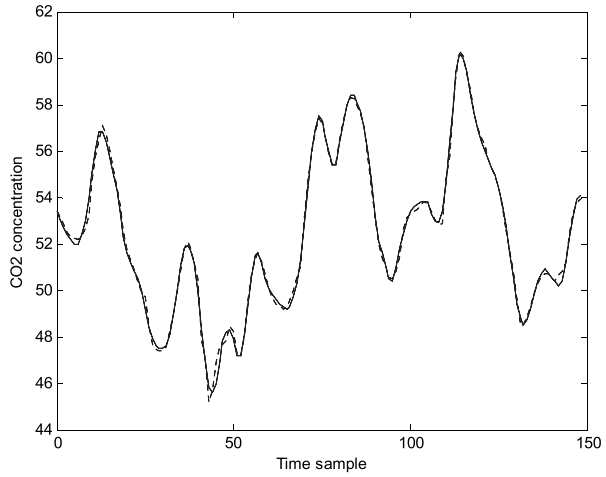

Fig. 7. Training in series–parallel configuration for model (34) of gas furnace by LP-SVR with projection operator wavelet kernel (46) and (47). Solid line: actual system output. Dotted line: model output.

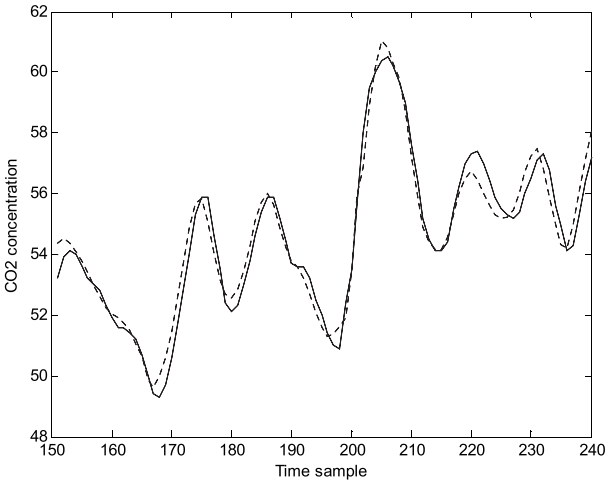

Fig. 8. Validation in parallel configuration for model (48) of gas furnace by LP-SVR with projection operator wavelet kernel (46) and (47). Solid line: actual system output. Dotted line: model output.

where $\hat{\boldsymbol{y}}_{n-1} = [\hat{y}_{n-1}, \hat{y}_{n-2}]$. The validation results are plotted in Fig. 8, and the corresponding RMSE is 0.5148. For confirming the superiority of the proposed kernel function, the model is also trained with other kernel functions (35)–(39) by the LP-SVR and the QP-SVR. Together with the model sparsity, the training RMSE and the validation RMSE are listed in Tables III and IV. It can be remarkably found from the tables that the validation accuracy obtained by using the multiscale projection operator wavelet kernel is dramatically improved and, in particular, the validation RMSE is even better than the training RMSEs from the Gaussian RBF kernel, polynomial kernel, and inverse multiquadric kernel. The corresponding plots can be found in [24].

Due to the ubiquity of transient characteristics and multiscale structures in nonlinear dynamics, the refinable kernel functions capable of taking account of local as well as global complexity are highly desired. Compared with the conventional single-scale kernel functions, the multiscale closed-form wavelet kernel functions display its main strength in capturing the localized temporal and frequency information of rapidly changing signals.

VI. Conclusion

The triumph of kernel methods largely depends on the capability of kernel functions. Confronted with the more and more challenging learning tasks, such as nonlinear dynamic systems identification, nonlinear time series prediction, and computer vision, the kernel machines are expected to be able to cope with the multiscale nature of complex systems.

Most of the kernel functions used in the literature, including the non-Mercer kernel in [48], are single-scale kernels. Criticized as template matchers, the commonly used translation-invariant kernels and rotation-invariant kernels may limit the performance that SV learning can achieve. In this paper, by leveraging the closed-form raised-cosine orthogonal wavelets to fulfill the finite expansion kernel, multiscale kernel learning was implemented in the framework of multiresolution analysis. In view of the geometric notion of the integral operator in function space, the developed finite multiscale expansion kernels are conceptualized as the multiscale projection operator wavelet kernel, thereby overcoming the limitation of the commonly used translation-invariant kernels and rotation-invariant kernels.

By focusing on control-oriented nonlinear dynamical systems modeling, the developed projector operator wavelet kernels are used to construct the composite kernel, and the sparsity inherited in linear programming SV learning ensures the lacunary kernel expansion representation for modeling nonlinear dynamic systems. Two examples have demonstrated the utility and effectiveness of the proposed projection operator wavelet kernel in representing nonlinear dynamic models in parallel configuration. The potential of the proposed kernel learning algorithm in hyperspectral image analysis, multiscale computer vision [60], and linear operator equations will be investigated further. On the theoretical aspect, the proposed multiscale projection operator wavelet kernels also shed light on the unexpected confluence of kernel regression and resolvent-type kernel-based nonuniform sampling [61], [62], which will enable us to explore the essence of SV selection in the LP-SVR from the perspective of modern sampling theory.

References

[1] Y. A. Abramovich and C. D. Aliprantis, Eds., *An Invitation to Operator Theory*. Providence, RI, USA: American Mathematical Society, 2002.

[2] P. G. Dodds, C. B. Huijsmans, and B. de Pagter, "Characterizations of conditional expectation-type operators," *Pacific J. Math.*, vol. 141, no. 1, pp. 55–77, 1990.

[3] A. Bobrowski, *Functional Analysis for Probability and Stochastic Processes*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[4] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[5] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Cambridge, MA, USA: MIT Press, 2001.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[7] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–202, Mar. 2001.

[8] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.

[9] M. Martínez-Ramón and C. Christodoulou, *Support Vector Machines for Antenna Array Processing and Electromagnetics*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2006.

[10] J. Krebs, "Support vector regression for the solution of linear integral equations," *Inverse Problems*, vol. 27, no. 6, pp. 1–23, 2011.

[11] J. Miura, "Support vector path planning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Beijing, China, Oct. 2006, pp. 2894–2899.

[12] S. De Marchi and R. Schaback, "Nonstandard kernels and their applications," *Dolomites Res. Notes Approx.*, vol. 2, no. 1, pp. 16–43, 2009.

[13] G. E. Fasshauer, "Positive definite kernels: Past, present and future," *Dolomites Res. Notes Approx.*, vol. 4, no. 1, pp. 21–63, 2011.

[14] R. Schaback and H. Wendland, "Kernel techniques: From machine learning to meshless methods," *Acta Numer.*, vol. 15, no. 1, pp. 543–639, 2006.

[15] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," in *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. Cambridge, MA, USA: MIT Press, 2007.

[16] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York, NY, USA: Academic, 2009.

[17] Z. Lu, J. Sun, and K. R. Butts, "Linear programming support vector regression with wavelet kernel: A new approach to nonlinear dynamical systems identification," *Math. Comput. Simul.*, vol. 79, no. 7, pp. 2051–2063, 2009.

[18] L. Zhang, W. Zhou, and L. Jiao, "Wavelet support vector machine," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 34–39, Feb. 2004.

[19] A. Widodo and B.-S. Yang, "Wavelet support vector machine for induction machine fault diagnosis based on transient current signal," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 307–316, 2008.

[20] Q. Wu, "The forecasting model based on wavelet v-support vector machine," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7604–7610, 2009.

[21] G. Y. Chen and W. F. Xie, "Pattern recognition with SVM and dual-tree complex wavelets," *Image Vis. Comput.*, vol. 25, no. 6, pp. 960–966, 2007.

[22] G. G. Walter and J. Zhang, "Orthonormal wavelets with simple closed-form expressions," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2248–2251, Aug. 1998.

[23] A. I. Zayed and G. G. Walter, "Wavelets in closed forms," in *Wavelet Transforms and Time-Frequency Signal Analysis*, L. Debnath, Ed. Boston, MA, USA: Birkhäuser, 2001, pp. 121–143.

[24] Z. Lu, J. Sun, and K. Butts, "Multiscale asymmetric orthogonal wavelet kernel for linear programming support vector learning and nonlinear dynamic systems identification," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 712–724, May 2014.

[25] R. Opfer, "Multiscale kernels," *Adv. Comput. Math.*, vol. 25, no. 4, pp. 357–380, 2006.

[26] A. Juditsky *et al.*, "Nonlinear black-box models in system identification: Mathematical foundations," *Automatica*, vol. 31, no. 2, pp. 1725–1750, 1995.

[27] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Berlin, Germany: Springer, 2001.

[28] O. Nelles, "On the identification with neural networks as series-parallel and parallel models," in *Proc. Int. Conf. Artif. Neural Netw.*, Paris, France, Oct. 1995, pp. 255–260.

[29] R. T. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, MA, USA: Birkhäuser, 1997.

[30] G. G. Walter and X. Shen, *Wavelets and Other Orthogonal Systems*. Boca Raton, FL, USA: Chapman & Hall, 2000.

[31] Y. Meyer and R. Coifman, *Wavelets: Calderón–Zygmund and Multilinear Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[32] A. N. Michel and C. J. Herget, *Applied Algebra and Functional Analysis*. Mineola, NY, USA: Dover, 1993.

[33] I. M. Glazman and J. I. Ljubič, *Finite-Dimensional Linear Analysis: A Systematic Presentation in Problem Form*. Mineola, NY, USA: Dover, 2006.

[34] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*. Mineola, NY, USA: Dover, 1993.

[35] W.-F. Zhang, D.-Q. Dai, and H. Yan, "Framelet kernels with applications to support vector regression and regularization networks," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1128–1144, Aug. 2010.

[36] K. Singla and J. L. Junkins, *Multi-Resolution Methods for Modeling and Control of Dynamical Systems*. Boca Raton, FL, USA: CRC Press, 2009.

[37] S. Xie, A. T. Lawniczak, S. Krishnan, and P. Liò, "Wavelet kernel principal component analysis in noisy multi-scale data classification," *ISRN Comput. Math.*, vol. 2012, no. 1, pp. 1–13, 2012.

[38] S. Xie, A. T. Lawniczak, and P. Liò, "Features extraction via wavelet kernel PCA for data classification," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Kittilä, Finland, Aug./Sep. 2010, pp. 438–443.

[39] A. Bonami, F. Soria, and G. Weiss, "Band-limited wavelets," *J. Geometric Anal.*, vol. 3, no. 6, pp. 543–578, 1993.

[40] J. A. Hogan and J. D. Lakey, *Duration and Bandwidth Limiting: Prolate Functions, Sampling, and Applications*. Boston, MA, USA: Birkhäuser, 2012.

[41] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.

[42] K. Urban, *Wavelet Methods for Elliptic Partial Differential Equations*. Oxford, U.K.: Oxford Univ. Press, 2009.

[43] M. Martínez-Ramón *et al.*, "Support vector machines for nonlinear kernel ARMA system identification," *IEEE Trans. Neural Netw.*, vol. 17, no. 16, pp. 1617–1622, Nov. 2006.

[44] Z. Lu, J. Sun, and K. Butts, "Linear programming SVM-ARMA$_{2K}$ with application in engine system identification," *IEEE Trans. Autom. Sci. Eng.*, vol. 8, no. 4, pp. 846–854, Oct. 2011.

[45] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[46] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[47] H. R. Zhang, X. D. Wang, C. J. Zhang, and X. S. Cai, "Robust identification of non-linear dynamic systems using support vector machine," *IEE Proc.-Sci., Meas. Technol.*, vol. 153, no. 3, pp. 125–129, May 2006.

[48] Z. Lu and J. Sun, "Non-Mercer hybrid kernel for linear programming support vector regression in nonlinear systems identification," *Appl. Soft Comput.*, vol. 9, no. 1, pp. 94–99, 2009.

[49] A. Gretton, A. Doucet, R. Herbrich, P. J. W. Rayner, and B. Schölkopf, "Support vector regression for black-box system identification," in *Proc. 11th IEEE Signal Process. Workshop Statist. Signal Process.*, Singapore, Aug. 2001, pp. 341–344.

[50] W. C. Chan, C. W. Chan, K. C. Cheung, and C. J. Harris, "On the modelling of nonlinear dynamic systems using support vector neural networks," *Eng. Appl. Artif. Intell.*, vol. 14, no. 2, pp. 105–113, Apr. 2001.

[51] T. Lin, B. G. Horne, P. Tiňo, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 7, no. 6, pp. 1329–1338, Nov. 1996.

[52] H. T. Siegelmann, B. G. Horne, and C. L. Giles, "Computational capabilities of recurrent NARX neural networks," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 27, no. 2, pp. 208–215, Apr. 1997.

[53] H. L. Wei and S. A. Billings, "Long term prediction of non-linear time series using multiresolution wavelet models," *Int. J. Control*, vol. 79, no. 6, pp. 569–580, 2006.

[54] K. Judd and M. Small, "Towards long-term prediction," *Phys. D, Nonlinear Phenomena*, vol. 136, nos. 1–2, pp. 31–44, 2000.

[55] G. Bontempi and S. B. Taieb, "Conditionally dependent strategies for multiple-step-ahead prediction in local learning," *Int. J. Forecasting*, vol. 27, no. 3, pp. 689–699, 2011.

[56] J. Sjöberg *et al.*, "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[57] D. Kukolj and E. Levi, "Identification of complex systems based on neural and Takagi–Sugeno fuzzy model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 272–282, Feb. 2004.

[58] H. Du and N. Zhang, "Application of evolving Takagi–Sugeno fuzzy model to nonlinear system identification," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 676–686, 2008.

[59] P. Wittek and C. L. Tan, "Compactly supported basis functions as support vector kernels for classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2039–2050, Oct. 2011.

[60] B. M. ter Haar Romeny, *Front-End Vision and Multi-Scale Image Analysis: Multi-Scale Computer Vision Theory and Applications, Written in Mathematica*. Dordrecht, The Netherlands: Springer, 2003.

[61] A. G. García and M. A. Hernández-Medina, "A general sampling theorem associated with differential operators," *J. Comput. Anal. Appl.*, vol. 1, no. 2, pp. 147–161, 1999.

[62] M. H. Annaby and A. I. Zayed, "On the use of Green's function in sampling theory," *J. Integral Equ. Appl.*, vol. 10, no. 2, pp. 117–139, 1998.

Zhao Lu (M'08–SM'15) received the M.S. degree in control theory and engineering from Nankai University, Tianjin, China, in 2000, and the Ph.D. degree in electrical engineering from the University of Houston, Houston, TX, USA, in 2004.

He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI, USA, and the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI, USA, respectively, from 2004 to 2006. Since 2007, he has been a Faculty Member with the College of Engineering, Tuskegee University, Tuskegee, AL, USA, where he is currently an Associate Professor with the Department of Electrical Engineering. His current research interests include machine learning, computational intelligence, and nonlinear control theory.

Jing Sun (M'89–SM'00–F'04) received the B.S. and M.S. degrees from the University of Science and Technology of China, Hefei, China, in 1982 and 1984, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA, in 1989.

She was an Assistant Professor with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI, USA, from 1989 to 1993. She joined the Ford Research Laboratory, Dearborn, MI, USA, in 1993, where she was with the Powertrain Control Systems Department. After spending almost ten years with the industry, she returned to academia and joined as a Faculty Member with the College of Engineering, University of Michigan, Ann Arbor, MI, USA, in 2003, where she is currently a Professor with the Department of Naval Architecture and Marine Engineering and the Department of Electrical Engineering and Computer Science. She holds over 30 U.S. patents and has co-authored the textbook entitled *Robust Adaptive Control*. Her current research interests include system and control theory and its applications to marine and automotive propulsion systems.

Prof. Sun was a recipient of the 2003 IEEE Control System Technology Award.

Kenneth Butts (M'10) received the B.E. degree in electrical engineering from Kettering University, Flint, MI, USA, the M.S. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA.

He is currently an Executive Engineer with the Powertrain and Regulatory Division, Toyota Motor Engineering and Manufacturing North America, Ann Arbor, where he is investigating methods to improve engine control development productivity. He has been involved in the field of automotive electronics and control since 1982, almost exclusively in research and advanced development of powertrain controls.